Performance Analysis of Automatic Lip Reading Based on Inter-Frame Filtering

Jinyoung Kim, Seungho Choi**,Seongmo Park**

*Dept. of Electronics, Computer and Information Engineering, Chonnam National University, Kwang-ju, Korea
**Dept. of Information & Communications Engineering, Dongshin University, Naju, Korea

kimjin@dsp.chonnam.ac.kr, smpark@chonnam.ac.kr, shchoi@white.dongshinu.ac.kr

Abstract

Automatic lip-reading has been focused as a complimentary method of automatic speech recognition in noisy environments. One of the most competitive lip-reading algorithms is the image transform based lip-reading (ITLR) algorithm. However, ITLR has severe performance degradation under illumination variations.

RASTA is a kind of inter-frame filtering method. It is used for rejecting stationary and convolutional noise in speech signal processing. In this paper, we apply RASTA approach to ITLR and analyze the performance of this method. We propose two merging techniques of pre-integration (PRE-I) and post-integration (POST-I). In PRE-I RASTA, inter-frame filtering is performed ahead of the image transform process. In POST-I, inter-frame filtering is done after the image transform process. We also compare the effectiveness of high-pass filtering and band-pass filtering as inter-frame filtering.

Experimental results show that pre-integration is very effective to reject illumination variances. And it is observed that high-pass filtering is enough to enhance the performance of lip-reading.

1. Introduction

Recently, researches on automatic lip-reading using the video sequence of the speaker's mouth have attracted significant interest. Automatic lip-reading under noisy environments is very effective in compensation for the decrease of speech recognition rate with an audio-only speech recognition (ASR) system [1]. The bimodal based on audio-visual information is an important part of the human-computer interface (HCI). We allow more weighting value to visual data than to audio one under a bad SNR but, on the contrary, more to audio data than to visual one under a clean SNR [2]. Under noisy circumstances, this bimodal approach has been a good alternative showing superior recognition rate to audio-only ASR system.

In this paper, we concentrate on the image transform based approach for automatic lip-reading (ALR) for bimodal speech recognition system. This approach is known to be superior to a lip-contour-based method for visual-only HMM recognition tasks. However, while the lip-contour based approach needs only several visual data, for example, outer, inner lip contour and lip width, the image-transform-based approach requires much larger visual feature vectors since it is based on the whole transformed image data of the speaker's mouth. Thus, for a fast algorithm, the necessity to reduce those data size has arisen.

To reduce the dimensionality of feature vectors, principal components analysis(PCA) has been suggested as a good method, which is based on linearly projecting the image space to a low dimensional feature space [3]. By the way, ITLR has the problem of robustness. Under varying illumination, the observed image sequences are suffered from rapid performance degradation. Illumination variation from the inconsistency of training and test conditions interferes the recognition process such as exact feature extraction. This interference causes a mismatching between the correct word and the related feature model and, after all, reduces the recognition rate. Our preliminary experiment in lip-reading system showed that even only a small amount of intensity variation caused large degradation of lip-reading performance [4].

To tackle those problems we propose the inter-frame filtering method, which is very similar with RASTA filtering in automatic speech recognition (ASR). According to reference [5], RASTA filtering is very successful in ASR under convolutional noisy environment. We propose two kinds of integration methods, pre-integration and post-integration. We examine usefulness of the inter-frame approach with our own lip-reading system.

In section 2, we briefly describe the algorithm for real-time automatic visual-only lip-reading system and mention about the necessity of the proposed method. Section 3 describes methods to diminish the illumination noise for the improved recognition rate. Finally, section 4 presents experimental results.

2. Baseline system : visual-only HMM-based lip-reading system

To develop a robust lip-reading algorithm, we implemented an automatic image transform based lip-reading system using HMM based word model. Figure 1 shows the overall block diagram of the implemented system based on the proposed algorithm. Given image sequence containing speaker's mouth, the overall process to extract the visual feature data consists of two sub-processes. One is ROI (region of interest) extraction process and the other is feature parameter extraction process.

2. 1 ROI extraction

Since lip-reading is based on the visual information of moving lip, extraction of appropriate interesting regions containing only moving lip area is important. ROI extraction from each image frame of given sequence is performed before feature extraction. As shown in figure 1, ROI extraction process consists of three steps; 1) gray-level transformation, 2) masking filtering and 3) binary-level transformation.

To find lip area efficiently, color image is first transformed into gray level image and then into binary-level image.

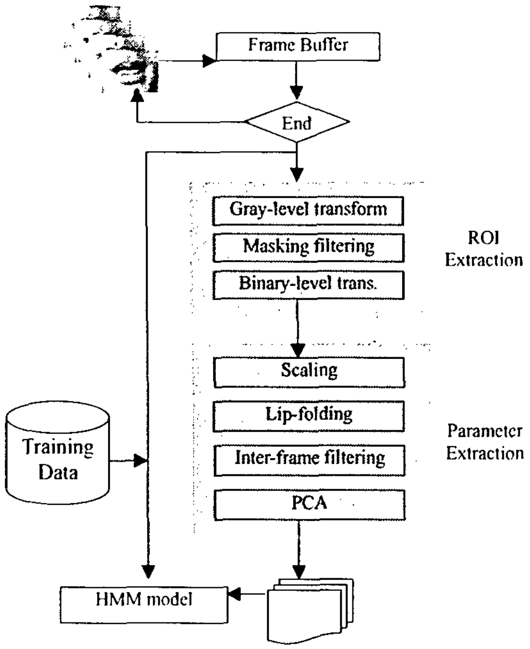


Figure 1. Block diagram of the proposed method for real-time visual-only HMM based lip-reading system

Both lip-ends of moving lip are extracted from this binary-level image by applying Y-projection and then X-projection. The vertical and horizontal center of speaker's mouth is obtained from these X, Y-projection. Then, the square pixel window of ROI is constructed around speaker's mouth. Since the lip width information of moving lip is important, we keep the width of ROI obtained at the first frame of each word to the last frame of that word. During the ROI extraction process, 'masking filter' is applied to diminish the unbalanced illumination of facial area from various lighting source.

2. 2 Feature extractions

To reduce the visual feature parameter size, each ROI is downsampled into a 16 x 16 pixel window for fast algorithm. This operation is necessary not only to reduce the feature data size but also to normalize the difference between each ROI size due to variations such as speaker's lip widths and the distances from camera.

To reduce the parameter size, dimensionality of visual feature vector, PCA (principal component analysis) is applied. PCA is known as a simple method to implement and to guarantee good performance in automatic lip-reading [6]. And, we use lip-folding technique before PCA process. Lip-folding is based on the symmetric property of lip along the vertical axis. Lip-folding makes 16 x 16 image size to half size of 8 x 16. The mean half-sized image needs smaller principal components to represent it than the original unfolded one. Additionally, the mean image compensates the illumination unbalance between the left lip area and the right lip area and, therefore, shows robustness under various lighting conditions[7].

2. 3 HMM based word recognition

For every video field, a static observation feature vector is acquired and those vectors obtained from the given video sequence are used for HMM based word modeling. Our automatic lip-reading system uses continuous density HMMs as a means of statistical pattern matching. The HMM observation probabilities are modeled as multi-dimensional Gaussian mixtures with diagonal covariance matrices. For the specific lip-reading recognition tasks considered in this paper, we use whole word, 3-6 state, left-to-right models with 3-8 mixtures per state. All HMM parameters are estimated by maximum likelihood Viterbi training.

3. Inter-frame filtering

One of ASR problems is the robustness. The performance of ASR is commonly worse in noisy environments. In general, noise is classified into additional and convolutional noise. RASTA filtering is one of methods used in ASR for preventing the degradation of ASR performance. RASTA is the abbreviation of 'relative spectral smoothing'. It was found that filtering time trajectories could compensate greatly for the effect of the convolutional noise induced by communication channel[5]. RASTA filtering is performed with bandpass filter. In RASTA filtering slow varying components, corresponding to the frequency characteristics of communication channel, are suppressed. The low-pass filtering helps to smooth some of the fast frame-to-frame spectral change present. The commonly used bandpass filter is as follows.

$$H(z) = 0.1z^4 \frac{2+z^{-1}-z^{-3}-2z^{-4}}{1-0.98z^{-1}} \qquad (1)$$

Based on these results, we discuss how inter-frame filtering is applied to lip-reading problems to enhance the performance of automatic lip-reading.

3.1 Integration of inter-frame filtering with lip-reading system

According to original work of Hermansky, RASTA filtering is applied to speech feature vector (SFV) sequence after obtaining SFVs. The RASTA filter is a kind of bandpass filter to reject slow and fast varying components. In our lip-reading system, feature extraction processing is PCA and the feature parameters are projection values of original image into most important axis. Thus, we can integrate inter-frame filtering after PCA in our lip-reading system, a simple imitation of ASR structure adopting RASTA filtering. We call this approach as post-integration (Post-I). Figure 2 shows the block diagram of Post-I method.

On the other hand, our AV database (DB) was recorded at various lighting conditions with illumination not regulated when visual DB was recorded. Thus, we may think that our AV DB was originally suffered from illumination noise. If the illumination noise was variant and dynamic, the result of PCA may include the influence of illumination noise. So, the *m* important axes would contain the components induced by illumination noise. This concept makes us change the order of PCA and inter-frame filtering. Figure 3 shows the second integration method of pre-integration (Pre-I).

3.2. Filters for inter-frame filtering

The band-pass filter used in ASR is shown in eq. (1). It is not impossible to use this filter for filtering image sequence. It's

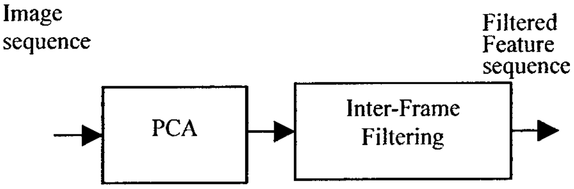


Figure 2. Post-integration method(Post-I).

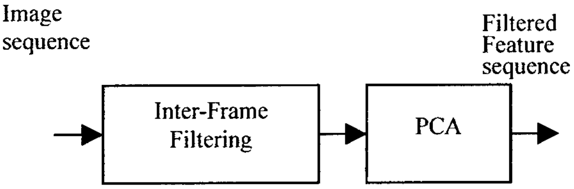


Figure 3. Pre-Integration method(Pre-I).

because the sampling frequency is very low in case of image capture operation compared with speech sampling. For speech signal 100 feature vectors per second is common. But, in our case, sampling frequency for image signal is 30Hz/second. So, we used very simple IIR filter for inter-frame filtering as follows.

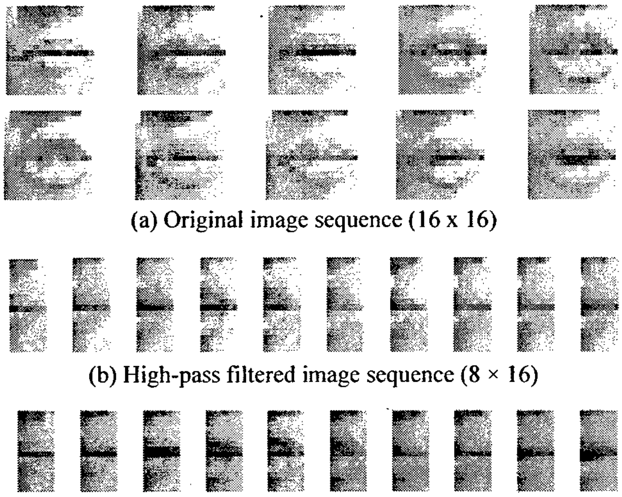
High-pass filter :

$$\begin{aligned} Y_t[n,m] &= 0.9858\cdot(X_t[n,m]-X_{t-1}[n,m]) \\ &\quad +0.9716\cdot Y_{t-1}[n,m] \end{aligned} \tag{2}$$

Low-pass filter :

$$\begin{aligned} Y_t[n,m] &= 0.8638\cdot(X_t[n,m]+X_{t-1}[n,m]) \\ &\quad +0.7257\cdot Y_{t-1}[n,m] \end{aligned} \tag{3}$$

Both filters are IIR(1,1) filters designed using MATLAB tool. Figure 4 shows the original image sequence and the filtered image sequences.



(a) Original image sequence (16 x 16)

(b) High-pass filtered image sequence (8 × 16)

(c) Band-pass filtered image sequence (8 × 16)

Figure 4. Inter-frame image filtering results

Table 1. Experimental environments.

Camara	SONY digital home video camera
Frame rate	30 frames/sec
Words	22 Korean words selected from the command menu for car navigation system
Training speakers	52 male speakers
Test speakers	18 male speakers different from training speakers
Recording condition	All recording are performed at different rooms at different time

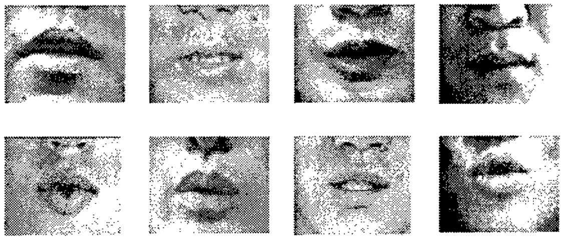


Figure 5. Some examples of our database recorded.

4. Experimental Environments and Results

4.1 Experimental environments

The experimental environment is shown in table 1. The database is composed of 22 Korean words spoken by 70 speakers. Figure 5 shows sample images of the AV database. As shown in the figure, our database recorded at different rooms and at different time, reveals illumination variations.

4.2 Experimental results

In this subsection, we describe the results of two proposed integration methods; Pre-I and Post-I, in the point of feature vector dimension and recognition results. Table 2 shows the dimension of features in Pre-I and Post-I integrations. From table 2, it is observed that post integration method is very effective in

Table 2. Comparison of feature dimensions in cases of Pre-I and Post-I

	Filter	PCA 90%	PCA 95%
Post-Integration	Bandpass	24	44
	Highpass	24	44
	NonFilter	24	44
Pre-Integration	Bandpass	6	14
	Highpass	6	14
	NonFilter	24	44

reduction of principal component numbers. The reason for this achievement could be that the pre-filtering rejects the influence of illumination noise before PCA process.

The other observation is that the low-pass filtering does not reduce the feature vector dimension. This result is not remarkable, for the sampling rate of image signal is much lower than that of speech signal. Anyway, using the post-integration, the feature vector dimension is reduced up to approximately 30%. The recognition results are shown in figure 6 and 7. From these two figures we can observe the following facts.

1) The post-integration doesn't improve the lip-reading performance. It makes the lip-reading performance worse. But the pre-integration enhance the recognition rate of the lip-reading system. This fact is the different point compared with the ASR.
2) The band-pass filtering, especially low-pass filtering is not decisive to increase the recognition rate. In other words, high-pass filtering is enough to the lip-reading system. As discussed above, it's because the sampling rate of video data is high when we consider the rate of lip movements in speaking.

It is obvious that pre-integration of inter-frame filtering is very effective in automatic lip reading. Pre-integration not only reduces the dimension of feature space but also improves the recognition rate of image-based lip-reading system.

5. Concluding Remarks

In general, lip-reading performance, especially image transform based lip-reading, is very sensitive to illumination variance. So, it is necessary to develop the robust version of lip reading to use automatic lip-reading in real service environments.

In this paper, we proposed inter-frame filtering approach as one of robust lip-reading methods and analyzed the performance of the proposed methods. From our experimental results we showed that pre-integration of inter-frame filtering enhanced lip-reading performances. The achievements are as follows.

1) Inter-frame filtering reduced feature vector dimension.
2) Inter-frame filtering improved the recognition rate of automatic lip reading.

In the future work, we will enlarge our AV database and study more robust methods so that automatic lip-reading can be used in. real environments

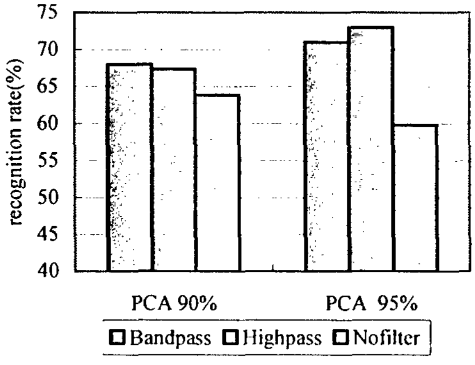


Figure 6. Recognition results of post-integration

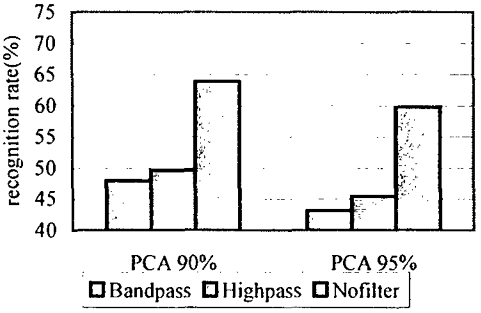


Figure 7. Recognition results of pre-integration

Acknowledgements

This work was supported partially by RRC HECS and ETRI.

References

[1] Meier U., Hürst W., Duchnowski P., " Adaptive Bimodal Sensor Fusion For Automatic Speechreading ", Proc. of ICASSP'96, Vol. 6. No. 2., pp. 833-836, 1996.

[2] Sharma R., Pavlovic V. I., Huang T. S., " Toward Multimodal Human-Computer Interface ", Proc. of IEEE, Vol. 86. No. 5., pp. 853-869, 1998.

[3] Sirovitch L. and Kirby M., "Low-Dimensional Procedure for the Characterization of Human Faces", J. Optical Soc. of Am., Vol. 2, pp.519-524, 1987.

[4] Kim J., Lee J. and Shirai K., "A Study on Various Factors Concerned with Lip-reading Performance at Dynamic Environment", J. of ICSP'01, pp.923-926, 2001.

[5] Hermansky H., Morgan M. and Bayya A., and Kohn P., "Compensation for the effect of the communication channel and auditory analysis of speech (RASTA-PLP)," Proc. Of Eurospeech'91, pp.1367-1371, 1991.

[6] Potamianos G., Graf H. P. and Cosatto E., "An Image Transform Approach for HMM Based Automatic Lipreading," Proc. of ICASSP '98, pp.173-177, 1998.

[7] J. Lee and J. Kim, "An efficient Lipreading Method Using the Symmetry of Lip," Proceedings of EuroSpeech2001, pp.1019-1022, 2001.

[8] Daubias P., Deleglise P., "Evaluation of an Automatically Obtained Shape and Apprearance Model For Automatic Audio Visual Speech Recognition," Proc. of Eurospeech2001, pp.1031-1034, 2001.

[9] Lièvin M. and Luthon F. "Lip Features Automatic Extraction", Proc. Of the 5th IEEE Int. Conf. On Image Processing. Chicago. Illinois, 1998.

[10] Uwe Meier, Rainer Stiefelhagen, Jie Yang, " Preprocessing of visual speech under real world conditions ", Interactive Systems Lab. European Tutorial & Research Workshop on Audio-Visual Speech Processing: Computational & Cognitive Science Approaches (AVSP 97).